DOI:
10.1111/papt.12171

Document Version
Peer reviewed version

Running head: *Self-criticism measures: systematic review*

Self-Criticism Self-Report Measures: Systematic Review

Alexandra V. Rose & Katharine A. Rimes*

King's College London, Institute of Psychiatry, Psychology and Neuroscience

*Author for Correspondence

Dept of Psychology, Henry Wellcome Building, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London SE5 8AF

Katharine.Rimes@kcl.ac.uk; Tel +4 (0)207 848 0330

To appear in Psychology and Psychotherapy

http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)2044-8341

Abstract

Purpose

Self-criticism is a transdiagnostic process that has been attracting research and clinical interest. The accurate measurement of this construct is therefore crucial, however, there are currently numerous measures of self-criticism and no guidelines about which to use in different contexts. This systematic review evaluated the measurement properties of self-report questionnaires of self-criticism.

Methods

OvidSP and Web of Science were used to search through multiple databases and an initial grey literature search was completed. Studies were included when the main focus was to evaluate the measurement properties of English version of scales or subscales that aimed to measure self-criticism in an adult population. Both the methodological quality of included studies and the specific measurement properties were evaluated; these ratings were then combined into a best evidence synthesis.

Results

Five scales and five subscales were identified, described in 16 papers. The scales were designed to measure different types of self-criticism including trait or repetitive self-criticism and self-criticism in response to difficult situations or as a mood regulation strategy. The majority of included studies were either rated as having poor methodological quality, or were given indeterminate or negative ratings for the measurement properties they reported.

Questionnaire content varied depending on how the authors conceptualised self-criticism. Issues were also highlighted in relation to the checklist used to rate methodological quality.

Conclusions

Tentative recommendations were made about two measures of self-criticism based on existing evidence; future research is required. Furthermore, questionnaire choice should be based on the type of self-criticism being assessed.

Practitioner Points

- Self-criticism has been associated with a range of clinical difficulties including depression and eating disorders, and is increasingly the focus of research, including treatment studies directly targeting self-criticism.
- Since different researchers have conceptualised self-criticism differently, a number of self-criticism self-report questionnaires have been developed that vary in terms of design, structure and content.
- This systematic review identified and evaluated the measurement properties of self-report questionnaires of self-criticism, and makes tentative recommendations about their use in clinical and research settings and areas for future research.

Key Words

Systematic review; Self-critical; Psychometric; Questionnaire; Assessment.

Introduction

Self-criticism is a self-evaluative process where individuals judge themselves in a negative or harsh way (Shahar et al., 2015a). Self-criticism has been described as a transdiagnostic process; it is associated with a wide range of problems including depression (Luyten et al., 2007), social anxiety (Shahar, Doron & Szepsenwol, 2015b) and eating disorders (Fennig et al., 2008). Since self-critical individuals have poorer outcomes after depression treatment (Marshall et al., 2008; Rector et al., 2000), research has begun to focus on treatments specifically targeting self-criticism (see Kannan & Levitt, 2013 for a review, as well as Shahar et al., 2012; 2015a for example treatment studies). Given the links between self-criticism and clinical disorders, it is vital to measure this construct using valid and reliable measures.

Researchers have conceptualised self-criticism differently leading to the development of a number of self-report questionnaires that differ in design and structure. Furthermore, as no 'gold standard' questionnaire exists, some researchers have attempted to measure self-criticism using a mixture of items from different measures (for example, Cox et al., 2004), or using questionnaires that were not originally developed to measure self-criticism, such as the Dysfunctional Assumptions Scale (DAS) (Weissman & Beck, 1978) or the original or revised versions of the Depressive Experiences Questionnaire (DEQ) (Blatt, D'Afflitti & Quinlan, 1976; Welkowitz, Lish & Bond, 1985; Bagby et al., 1994; Viglione et al., 1995; Santor, Zuroff & Fielding, 1997). Although the DEQ contains a factor called 'self-criticism', this factor aims to measure 'introjective depression', rather than the construct of self-criticism.

Having multiple measures of self-criticism gives practitioners choice regarding which questionnaire to use in different contexts. On the other hand, it could be difficult to decide

which is most appropriate, particularly when it is unclear about which questionnaires are of adequate psychometric quality (de Boer et al., 2004), or when attempting to use a questionnaire that has been validated in a non-clinical population with different patient groups. Furthermore, if different measures are used, the comparison of results between studies is very difficult. Finally, using questionnaires that were not originally designed to measure self-criticism may lead to uncertainty about the interpretation of their findings and conclusions.

Objectives

This systematic review evaluated the measurement properties of self-report questionnaires of self-criticism. The characteristics, for example, length and content area, and psychometric properties are reviewed and recommendations about their use are made.

An initial scope of the literature highlighted that there are currently different research groups interested in self-criticism, each defining this construct differently. This review defined self-criticism as a self-evaluative process involving negative or judgemental thoughts about the self. This review was concerned with measures of self-criticism in general rather than a single specific type of self-criticism, or self-criticism focused on circumscribed areas such as one's appearance or academic ability.

Method

Selection Criteria

Papers were included if they were full text original articles published in English in a peer reviewed journal, where the main focus was to describe the development or evaluation of the measurement properties of an English-language self-report questionnaire (or subscale) that aimed to measure self-criticism (or a synonym of self-criticism) in an adult population

(both clinical and non-clinical populations). The measure or subscale needed to focus on self-criticism in a general way across different life domains (rather than focusing on self-criticism about one particular activity, for example, sport). Papers were excluded if they described broader measures of negative automatic thoughts, or if they only focused on one specific type of self-criticism, such as self-blame. Papers were also excluded if the items of the questionnaire could not be extracted or located (after internet searches, inter-library loan requests & authors directly contacted).

Search Strategy

OvidSP and Web of Science (WoS) were used to search through a number of databases. In WoS, the Core collection & Medline were both selected, excluding case reports, and refined by English language. In OvidSP, PsycINFO and Embase Classic+Embase (1947 to date of search) were selected, with English language added as a limit. The initial search took place in June 2015 and it was updated in October 2017.

To limit publication bias, an initial scope of the Grey literature was completed in June 2015 (Mahood, Van Eerd & Irvin, 2014), using the following databases: The Open Grey Repository, DART Europe E-theses portal, EThOS, Open Access Theses & Dissertations and Global ETD Search. The authors were therefore fairly certain no relevant unpublished papers were being excluded; thus, only published articles were included in the review.

Search Terms

Search terms were: “self critic*” OR “inner critic*” OR “negative think*” OR “negative self statements” OR “self judg*” OR “self attitude*” OR “attitude* toward self”

AND

Psychometric* OR reliab* OR valid* OR reproducib* OR construct* OR develop* OR creat* OR assess*.

Some broader search terms were included as this review included subscales of self-criticism and authors may not have included all of the subscales names in the abstract. Several of the psychometric search terms were chosen from a previously developed search filter (Terwee et al., 2009). Furthermore, COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) recommends not using a search term for the "type of measurement instrument", for example, questionnaire, inventory, scale, etc, to reduce the risk of inappropriately excluding a paper (COSMIN, 2016).

Selection Process

The titles and abstracts of articles were screened and, after this, the full-text of the remaining articles were assessed for eligibility. Both of these stages were completed by the first author and, where unclear, the inclusion / exclusion of studies was discussed with the second author. Reference checking and citation tracking using OvidSP and Google Scholar were carried out on included studies.

Data Extraction & Quality Assessment

The questionnaire characteristics, study characteristics and evaluated measurement properties were extracted from included studies. The quality assessment was completed in three stages. To ensure that the included articles met the inclusion / exclusion criteria and the quality assessment was accurate, five articles were double rated by two independent reviewers, the main author and another trainee clinical psychologist who was familiar with COSMIN. The strength of agreement between reviewers was 'very good' [k= 0.88, p<0.0005] (Altman, 1999).

Step One – Assessment of The Methodological Quality of Studies

The methodological quality of the included studies was assessed using COSMIN (Mokkink et al., 2010a), currently the only specific tool focused on methodological quality assessment. The following domains are covered by COSMIN: internal consistency, reliability, measurement error, content validity, construct validity (divided into structural validity, hypothesis testing and cross-cultural validity), criterion validity and responsiveness. When a study assessed the same measurement property in multiple participant groups or different measurement properties using different participant groups, the relevant COSMIN boxes were completed more than once for that study.

For each study or sample, the methodological quality for a particular measurement property was rated by a series of items on a 4-point nominal rating scale: poor, fair, good, and excellent (Terwee et al., 2012). For each measurement property, COSMIN produces an overall score by the lowest rating of any item, i.e. the "worst score counts".

Step Two – Quality Assessment of Instruments

The assessment of the quality of each questionnaire was completed using the criteria proposed by Terwee et al. (2007). No criterion is provided for structural validity. Instead, for exploratory factor analyses, Schellingerhout et al. (2012)'s criteria was used, and for confirmatory factor analyses, a criteria was devised by the authors after consultation with two trainee clinical psychologists familiar with systematic reviews of measurement properties (see Table 1).

Step Three – Best Evidence Synthesis

In order to summarise the evidence of the measurement properties for each questionnaire, a best evidence synthesis was completed. This was done by combining the

results of different studies, taking account of the number and methodological quality of the studies (using COSMIN), the results of the measurement properties that were evaluated (see Table 1), as well as the consistency of results across studies. Each questionnaire was given an overall rating using a criteria similar to that proposed by the Cochrane Back Review Group (see Furlan et al., 2009; van Tulder et al., 2003), adapted from Schellingerhout et al. (2012) and has been used by a recent systematic review (Heinl et al., 2016).

[Insert Table 1 here]

Measurement Properties in This Study

The terminology and definitions of measurement properties used in this review are taken from Mokkink et al. (2010b). The measurement properties are divided into 3 domains: reliability, validity and responsiveness. Reliability consists of internal consistency, reliability and measurement error. Validity contains content validity (including face validity), construct validity and criterion validity. The term responsiveness is used for both the domain and measurement property.

Since there is no 'gold standard' measure of self-criticism, criterion validity was not assessed. Of the included studies, no information was provided for responsiveness and measurement error. In relation to the COSMIN interpretability box, information was only provided about how missing items were handled, and scores (i.e. means and standard deviations), meaning that the Terwee et al. (2007) properties 'floor and ceiling effects' and 'interpretability' were not completed.

Results

Selection of Studies

The PRISMA flowchart is displayed in Figure 1. In the database search, 4414 papers were found. Through the initial grey literature search, an unpublished version of the Self-Critical Rumination Scale was found and through contact with the author the (very recently) published paper was included for screening (Smart, Peters & Baer, 2015). Removing duplicates left 2693 papers that were screened and 2557 papers were excluded. The full text of 136 papers were reviewed and 125 of these were excluded. Of the 11 included papers, both reference checking and citation tracking resulted in the addition of 1 paper each (2 in total). Another additional paper was found through manual searching from a study that was screened at the full text stage. Although the study was excluded, through an internet search and contact with the author, the Temperament & Personality Questionnaire's original development paper was found and included (Parker et al., 2006).

The updated search in October 2017 resulted in a total of 569 additional papers which were screened. Two papers were taken to full text; 1 was excluded as it focused on questionnaire not measuring self-criticism, and the other was excluded as it was a review paper. However, the latter paper referenced an original article examining the factor structure of the Self-Compassion Scale (Neff, Whittaker & Karl, 2017), which was then included. Citation tracking also resulted in an additional paper being found (Brenner et al., 2017). Thus, a total of 16 papers were included in the review.

[Insert Figure 1 here]

Questionnaires Found in Search

This systematic review identified five questionnaires solely measuring self-criticism and five subscales measuring self-criticism. In the next section, the questionnaires are described in terms of the way in which they conceptualised self-criticism and then the results of the best evidence synthesis are discussed in relation to each measurement property (see also Table 2 for a summary of the best evidence synthesis). Please see additional Tables for the questionnaire characteristics (Table 3), the characteristics of the included studies (Table 4), and the methodological quality and measurement properties for internal consistency, reliability, content validity and structural validity for every study included (Table 5). Construct validity (measured through hypothesis testing) for every study included is displayed in Table 6, which includes correlation coefficients and between group comparisons. Only correlations were extracted for constructs that were deemed to be the most relevant for self-criticism research, including self-criticism using a different questionnaire, self-esteem, self-compassion, depression, general measures of anxiety, perfectionism, shame and rumination.

[Insert Tables 2 - 6 here]

Questionnaire Conceptualisation of Self-Criticism and Item Content

The following categories were identified regarding conceptualisations of self-criticism that underpinned the questionnaires: self-criticism as a trait, self-criticism in response to difficult situations, self-criticism as a mood regulation strategy and measures of repetitive self-criticism. These are summarised briefly below, together with information about item content where appropriate.

Self-criticism as a Trait

Two scales and three subscales focused on trait self-criticism. Both the Self-Critical Cognition Scale (SCCS) (Ishiyama & Munsun, 1993) and Levels of Self-Criticism Scale (LSCS) (Thompson & Zuroff, 2004) defined self-criticism as a dispositional tendency or broad personality construct. The SCCS's development paper describes the "self-defeating cognitive patterns and negative informational contents" found in clinical populations (Ishiyama & Munsun, 1993, P. 147). The LSCS is based on a theoretical model of personality development whereby personality proceeds through the simultaneous development of 'relatedness' and 'self-definition' and problems with these developmental lines may lead to 'dysfunctional' personality characteristics of 'dependency' and 'self-criticism' (Blatt, 1974). The LSCS aims to measure two developmental levels of self-criticism; self-criticism based on external standards and self-criticism related to internal standards at the other (Thompson & Zuroff, 2004).

Three questionnaires contain subscales that relate to self-criticism as conceptualised in relation to depression. The Temperament & Personality Questionnaire (TPQ) (Parker et al., 2006) Self-Criticism subscale defines self-criticism as a personality construct predisposing individuals to depression. The authors suggest that whilst the TPQ overlaps with the Five Factor Model of personality (McCrae & Costa, 1987), they describe the TPQ as a flexible measure capturing both higher-order and lower-order components that potentially predispose an individual to depression, and thus the inclusion of self-criticism as a lower construct (Parker et al., 2006).

In the Attitudes Towards Self Scale (ATSS) development paper, self-criticism is conceptualised as one of three potential vulnerabilities to depression (Carver & Ganellen,

1983). The ATSS measures the depressive tendency to be 'self-punitive' and includes three subscales which measure self-criticism, having high standards and overgeneralising specific failures as reflecting an individual's global self-worth. Self-criticism is described as an intolerance to the perceived discrepancy between the real and the desired self.

In terms of the actual items included in these questionnaires, the SCCS, ATSS and ATSSR (Carver et al., 1988) Self-Criticism subscales all included broader items about aspects other than self-criticism, such as the inability to keep a balanced perspective and the exaggeration of negative aspects of oneself (SCCS), or about reactions to failure (ATSS & ATSSR). Furthermore, the LSCS did not actually mention the term self-criticism. Additionally, although the subscale items of the 109-item version of the TPQ appeared to have a more specific focus on self-criticism or being tough on oneself, the TPQ research team have cautioned the use of this version due to confusion over scoring (R. Graham, personal communication, January 15, 2016).

Self-criticism in response to difficult situations

One scale and one subscale focused on self-criticism when things go wrong for someone, or in difficult times. Firstly, the Forms of Self-Criticizing/Attacking Reassuring Scale (FSCRS) (Gilbert et al., 2004) is based on Gilbert's social mentality theory, which proposes that competencies for co-ordinating roles with others are also used in evaluating the self (Gilbert, 2005). Gilbert suggests that there are different 'forms' of self-criticism, specifically that feeling inadequate can be separated from feelings of self-hatred, and that these can be differentiated by an individual's ability to self-reassure. The FSCRS included items about self-criticism and other negative feelings about oneself in relation to failure such as disappointment, inadequacy and disgust. Gilbert has also developed the Functions of Self-

Criticism/Attacking Scale, however, this was not included in this review as it focused on 'why' people engage in self-critical thinking, rather than measuring self-criticism per se (Gilbert et al., 2004).

The Self-Compassion Scale (SCS) (Neff, 2003) contains a Self-Judgement subscale; self-judgement is considered to be a synonym of self-criticism in this review. Neff (2003) conceptualises self-judgement as being at the opposite end of a continuum from self-kindness and this subscale is therefore reverse-scored when used to contribute to an overall measure of self-compassion. Neff (2003) describes self-kindness as seen as one of three basic components of self-compassion. The items of the Self-Judgement subscale focus on being disapproving and intolerant about one's flaws and being tough on oneself at times of suffering.

Self-criticism as a mood regulation strategy

The authors of the Inventory of Cognitive Affective Regulation Strategies (ICARUS) (Kamholz et al., 2006) describe how all psychological distress is related to affect dysregulation and that the Self-Criticism/'Self-Blame' subscale is considered to be one of many cognitive strategies that an individual might use when experiencing negative affect (Kamholz et al., 2006). They highlight the importance of using cognitive strategies to attempt to reduce emotional distress. The items focus on thoughts about faults and mistakes, as well as broader negative thinking in response to difficult emotions.

Repetitive self-criticism

Two measures of repetitive self-criticism were identified. Firstly, the Habitual Index of Negative Thinking (HINT) (Verplanken et al., 2007) was the only measure that focused solely on the process of negative self-thinking (considered to be a synonym of self-criticism)

as a mental habit, as opposed to focusing on the content. They define a habit as "…learned sequences of acts that have become automatic responses to specific cues, and are functional in obtaining certain goals or end states" (Verplanken & Aarts, 1999, P. 104). The items therefore focused on different aspects of the concept of a habit, such a frequency, lack of conscious intent and lack of awareness of initiation (Verplanken et al., 2007).

The authors of the Self-Critical Rumination Scale (SCRS) defined self-criticism as a "form of negative thinking that devalues the self", as opposed to a more global personality dimension (Smart et al., 2015, P. 2). They explain that the scale also focuses on the ruminative qualities of thinking in terms of being "frequent, prolonged, repetitive and difficult to control". Items therefore focused on both the process of self-criticism in terms of frequency and repetitiveness and as well as its content, including feeling ashamed of oneself.

Results of Best Evidence Synthesis

The results of the best evidence synthesis are provided in Table 2 and summarised briefly in the sections below.

Content validity

For content validity, all the questionnaires, apart from the Self-Criticism Rumination Scale (SCRS) (Smart et al., 2015), were rated as poor for the methodological quality of the studies and were given indeterminate ratings for the measurement property, resulting in 'weak' evidence in the best evidence synthesis (Table 2). This was because COSMIN states that the target population needs to be defined and involved in the item development or selection (and higher ratings are given depending on the sample size of the target population used); questionnaires other than the SCRS did not do this. The content validity of the SCRS was assessed using both undergraduates and out-patients at a mental health clinic (Smart et

al., 2015). As enough information was provided to assume that these were the target populations, the methodological quality was rated as excellent and it was given a positive rating for content validity.

Internal Consistency

The best evidence synthesis (see Table 2) gave highest internal consistency ratings to the FSCRS, which had 'moderate' positive evidence as its internal consistency was assessed in studies in which the methodological quality was rated as fair (Gilbert et al., 2004; Kupeli et al., 2013 & Baião et al., 2015). The SCCS (Ishiyama & Munsun, 1993), LSCS (Thompson & Zuroff, 2004) and SCRS (Smart et al., 2015) had 'limited' positive evidence as this measurement property had been investigated in studies of fair methodological quality. Both the SCS (Neff, 2003; Williams et al., 2014; Neff et al., 2017; Brenner et al., 2017) and ICARUS (Kamholz et al., 2006) had 'conflicting findings' for internal consistency due to inconsistencies across studies. There were issues with the Cronbach alpha calculations for the other questionnaires resulting in a 'limited' indeterminate rating for the TPQ (Parker et al., 2006), 'limited' negative ratings for the ATSS (Carver & Ganellen, 1983) and ATSSR (Carver et al., 1988) and a 'weak' rating for the HINT (Verplanken, 2006; Verplanken et al., 2007).

Test-Retest Reliability

The SCCS (Ishiyama & Munsun, 1993), SCS (Neff, 2003), SCRS (Smart et al., 2015) and ATSSR (Carver et al., 1988) were given 'limited' indeterminate ratings for test-retest reliability in the best evidence synthesis (see Table 2) because it was assessed using a statistical test other than what COSMIN recommends (i.e. Intraclass Correlation Coefficients were not used). However, of note, the test-retest correlation of the SCCS, SCS and SCRS

were high, whilst the test-retest correlation for the ATSSR was low. For the HINT (Verplanken et al., 2007), TPQ (Parker et al., 2006) and ICARUS (Kamholz et al., 2006), 'weak' ratings were given due to issues around participant stability between time points or sample size. The test re-test reliability of the LSCS (Thompson & Zuroff, 2004), ATSS (Carver & Ganellen, 1983) and FSCRS (Gilbert et al., 2004) has not been studied.

Structural Validity

Both the FSCRS (Gilbert et al., 2004; Kupeli et al., 2013 & Baião et al., 2015) and SCRS (Smart et al., 2015) were given 'moderate' positive ratings for structural validity because it had been assessed in multiple studies of fair methodological quality, whilst the SCCS (Ishiyama & Munsun, 1993), TPQ (Parker et al., 2006) and ATSSR (Carver et al., 1988) were given 'limited' positive ratings as this measurement property was only assessed in their development papers (see Table 2). Lower ratings were given for the LSCS (Thompson & Zuroff, 2004) as the variance explained by the final factors was not mentioned and for the ATSS (Carver et al., 1988) because the final factors only explained 40% of the variance. As with internal consistency, the SCS (Neff, 2003; Williams et al., 2014 & Neff et al., 2017; Brenner et al., 2017) and ICARUS (Kamholz et al., 2006) both had 'conflicting findings' for structural validity due to inconsistencies between studies. Structural validity for the HINT (Verplanken et al., 2007) has not yet been investigated.

Construct Validity (Hypothesis Testing)

In the best evidence synthesis, construct validity was assessed through hypothesis testing; higher ratings are given by COSMIN where hypotheses about the relationship between self-criticism and other variables were devised a priori. COSMIN suggests that when

the expected results are unclear or not found, this could suggest that the questionnaire is not measuring the construct it is aiming to.

For hypothesis testing, the SCS (Neff, 2003) and HINT (Verplanken, 2006; Verplanken et al., 2007) received 'moderate' positive ratings and the SCRS (Smart et al., 2015) received 'limited' positive evidence because results were in line with specific hypotheses (although, of note, correlations for the SCS subscales were not presented separately); see Table 2. For the FSCRS, although two studies had positive ratings (Gilbert et al., 2004; Kupeli et al., 2013), as the third had unclear hypotheses (Baião et al., 2015), this resulted in 'conflicting findings'. Other lower ratings were given to other questionnaires due to issues with the hypotheses made in studies which were either vague or not stipulated at all (and due to the study findings, it was unclear what was expected). The construct validity of the TPQ has not been studied (Parker et al., 2006).

Discussion

The aim of this systematic review was to identify and evaluate the measurement properties of self-report measures of self-criticism. The main theme that emerged was that most studies were either rated as having poor methodological quality, or were given indeterminate or negative ratings for the measurement properties they studied.

As well as this, a key theme emerged around the conceptualisation of self-criticism; the construct was defined differently by authors, leading to questionnaires with different content and structure. At times, the way self-criticism was defined was very broad or imprecise. Having an unclear definition of self-criticism could impact on item development and consequently measurement properties. Also, for some measures there appeared to be a disparity between what a questionnaire aimed to measure, and the actual items used. As the

focus of this systematic review was on questionnaires that "aimed" to measure self-criticism, the individual items were not formally evaluated. Nevertheless, as highlighted in the results, some items could be construed as measuring different affect or reactions to failure, or other distinct but overlapping constructs such as perfectionism, shame or self-esteem.

The differences between definitions, and the inclusion of items measuring broader or related constructs, appeared to reflect the different theoretical positions of the questionnaire's authors. As well as this, it is important to recognise this paper's own definition of self-criticism, as a broad self-evaluative process that includes negative self-content, and how this ultimately guided the inclusion / exclusion criteria of the current review. It was hoped that by defining self-criticism in a general way, it would capture questionnaires with the most potential to be used for different purposes in a variety of settings.

This review also attempted to divide the questionnaires according to different categories, including: self-criticism as a trait, self-criticism in response to difficult situations, self-criticism as a mood regulation strategy and measures of repetitive self-criticism. Even within these subgroups, authors still varied in their conceptualisations of self-criticism. For example, in the 'trait' category, the LSCS focused on self-criticism in the context of a psychodynamic developmental view of personality (Thompson & Zuroff, 2004), whilst the TPQ (Parker et al., 2006) considered self-criticism to be a lower order factor in the wider context of a model of personality overlapping with the Five Factor model (McCrae & Costa, 1987). The LSCS was linked to research about Blatt's depression vulnerability theory and the widely used Depressive Experiences Questionnaire (DEQ) (Blatt, D'Afflitti & Quinlan, 1976). It's important to emphasise that the DEQ aims to measure 'introjective depression' rather than the construct self-criticism. Thus, the DEQ's 'self-criticism' factor contains items

that reflect a range of constructs such as guilt and emptiness and being unable to assume responsibility. The TPQ, ATTS (Carver & Ganellen, 1983) and ATTSR (Carver et al., 1988) also defined self-criticism as a personality dimension linked it to a predisposition to depression. The relationship between self-criticism and depression has been well established (Luyten et al., 2007), making these questionnaires potentially useful within this context.

This review identified one scale and one subscale that focused on self-criticism in response to difficult situations. Firstly, the Forms of Self-Criticizing/Attacking Reassuring Scale (FSCRS) is unique in that it conceptualises self-criticism as a multi-dimensional construct with different forms, distinguishing between self-criticism about inadequacy from self-hatred (Gilbert et al., 2004). It also included an additional subscale measuring self-supportive thinking after things have gone wrong for an individual.

The FSCRS could be a useful outcome measure in research to assess treatment protocols (e.g. Gilbert, 2009) to reduce thoughts about inadequacy and self-hatred and improve an individual's ability to self-soothe. Furthermore, although not assessed in this review, the Functions of Self-Criticism/Attacking Scale may be a useful additional scale to consider why an individual engages in self-critical thinking in more detail (Gilbert et al., 2004).

The Self-Compassion Scale (SCS) conceptualised self-judgement as a negative component of self-compassion, placing self-judgement (or self-criticism) on a continuum with self-kindness (Neff, 2003). This raises questions about the relationship between self-criticism, self-kindness and self-compassion. The SCS is a widely used measure of self-compassion, however, this review highlighted that it requires further research to establish

whether the subscales should be used as standalone measures of particular constructs such as self-criticism / self-kindness.

This review identified two measures of repetitive self-criticism; the Self-Critical Rumination Scale (SCRS) (Smart et al., 2015) and the Habitual Index of Negative Thinking (HINT) (Verplanken et al., 2007). The SCRS was the only questionnaire to receive a positive rating for content validity, and had consistently positive ratings for its other psychometric properties assessed in this review. The SCRS also captures the ruminative and repetitive nature of self-critical thinking, a process that appears to be important in clinical populations (Nolen-Hoeksema, Wisco & Lyubomirsky, 2008).

The HINT was the only measure that focused on the habitual process of negative self-thinking, measuring different features of the concept of a habit. The HINT could be a useful outcome measure where researchers do not wish to use a measure that specifies the content of the self-critical thoughts. However, the methodological quality of the studies was poor, so further research is required to assess the psychometric properties in high quality studies.

Considering the different conceptualisations of self-criticism, self-criticism could be a multi-layered construct, with questionnaires measuring different aspects. For example, a self-critical individual may be predisposed to think in this way due to an underlying personality dimension that developed due to various negative early life experiences. This individual may then respond to further negative experiences with self-criticism to regulate their affect. This could further lead to a habitual pattern of responding to difficult situations in this way. To consider these ideas further, it maybe that a further substantive review and discussion is required about how to define self-criticism and to provide further clarity and reach a

consensus about this psychological construct. Until this time, the current review gives an opportunity to consider the strengths and potential limitations about included measures.

Assessing the Methodological Quality of Included Studies

This systematic review used COSMIN to assess the methodological quality of the included studies. COSMIN uses a "worst score counts" method whereby an overall score is determined by the lowest rating of any item. A number of themes emerged in regards common areas on which studies received lower ratings.

Firstly, apart from the SCRS, all measures were given a poor methodological quality rating for content validity. The COSMIN item where all studies fell down on was: "was there an assessment of whether all items are relevant for the study population?" Items were commonly developed by the authors. This is important as they would be considered 'experts' in their field, however, according to COSMIN, it is also crucial to define a target population and use these individuals to assess the included items. The majority of studies also used undergraduates, and it cannot be assumed that this was the intended target population of the questionnaire. This is particularly important given that researchers may wish to measure self-criticism in a wide range of non-clinical and clinical populations.

Secondly, for test-retest reliability only one study (Parker et al., 2006) explicitly stated that Intraclass Correlation Coefficients (ICCs) were calculated. COSMIN states that ICCs are the preferred statistical method for test-retest reliability with continuous scores as Pearson's and Spearman's correlation coefficients do not take account of systematic error (Mokkink et al., 2012). Because ICCs were rarely cited, the test-retest reliability correlation has been mentioned for each scale, i.e. whether it was high or low. This seemed appropriate because, although scales received the same COSMIN rating, some had test-retest reliability

correlations that were less than 0.70, which is widely accepted as being an unacceptable test-retest reliability value (Test-Retest Coefficient, 2016).

A possible limitation of COSMIN is that it does not have criteria to evaluate the face validity of each questionnaire, defined as the degree to which items of a questionnaire look as though they are an adequate reflection of the construct to be measured (Mokkink et al., 2010b). In the COSMIN manual it states that because face validity involves subjective judgement no criteria has been developed (Mokkink et al., 2012). Thus, in COSMIN, content validity focuses on whether the original authors assessed the relevance and comprehensiveness of items, rather than asking the COSMIN user to assess this. It is therefore possible that a questionnaire could be given positive ratings for content validity in terms of both the study's methodological quality and its measurement properties, but there could still be questions about whether the items actually measure self-criticism. Therefore, researchers and clinicians selecting a measure are strongly urged to check that the scale probe question, items and response ratings assess the construct of interest to them, rather than focusing purely on the findings from the best evidence synthesis. It would be helpful for future research to focus on the development of a set of criteria to formally assess face validity. This may be particularly pertinent for research areas such as self-criticism where there is no universally agreed definition of this construct.

Recommendations

Tentative recommendations are given based on the current level of evidence. However, due to the different conceptualisations of self-criticism, the questionnaire of choice will ultimately depend on the particular research approach or question. Where

recommendations are made to conduct 'high quality' research, the authors are referring to high methodological quality as operationalised by COSMIN.

Since the SCRS had consistent positive ratings, this review would recommend its use in future research if the focus is on frequent or repetitive self-critical thinking. This review would also recommend the FSCRS for researchers or clinicians wishing to assess self-criticism in response to things going wrong, particularly if they also wanted to assess self-hatred separately and / or self-reassurance. It would be important for future studies to focus on its test-retest reliability as this has not yet been assessed.

Due to the limited positive evidence for questionnaires that define self-criticism as a trait, it is recommended that future research either conduct high quality studies focused on their measurement properties or develop and evaluate alternative measures. In regard to subscales of self-criticism, because of the general limited positive, conflicting or weak evidence found, this review cannot make strong recommendations about their use to measure self-criticism.

Limitations of the review

The current review was not pre-registered which can be useful to avoid unnecessary duplication and also allows for a comparison between intended protocols and the final method used. Selection bias could have come from the fact that only English versions of self-report measures were included. However, the inclusion of translated versions could have resulted in inconsistent findings regarding the measurement properties of the same questionnaire, thus, it has been suggested that separate systematic reviews are conducted for translated versions of measures (Schellingerhout et al., 2012). Other areas of potential bias could have come from the fact that only one author carried out the initial search and

screening. On the other hand, when unclear, the inclusion / exclusion of studies was discussed with the second author, and 5/15 papers were doubled rated by an additional researcher. Finally, as with all reviews, there is a time lag between the literature search and publication. It is therefore recommended that the review is updated in 5 years time to allow for further research to be published.

Conclusions

Valid and reliable measures of self-criticism are needed to ensure that this construct is adequately measured across settings. This systematic review evaluated the measurement properties of scales and subscales measuring self-criticism, as well as assessing the methodological quality of included studies. Five scales and five subscales were found which were designed to measure four main different types of self-critical thinking. The content of the questionnaires therefore varied depending on the theoretical orientation of the authors. Tentative recommendations were made about the use of the SCRS and the FSCRS on the basis of existing evidence, however, further research is needed into these and some of the other scales. Due to differences between the precise focus of measures, the final decision about which questionnaire to use will ultimately depend on the goals of the researcher or clinician.

References

Altman, D. G. (1999). *Practical statistics for medical research*. New York, NY: Chapman & Hall/CRC Press.

Baião, R., Gilbert, P., McEwan, K., & Carvalho, S. (2015). Forms of Self-Criticising/Attacking & Self-Reassuring Scale: Psychometric properties and

normative study. *Psychology and Psychotherapy: Theory, Research and Practice*, *88*(4), 438-452. DOI: 10.1111/papt.12049

Bagby, R. M., Parker, J. D., Joffe, R. T., & Buis, T. (1994). Reconstruction and validation of the Depressive Experiences Questionnaire. *Assessment,1*(1), 59-68. DOI: 10.1177/1073191194001001009

Blatt, S. J. (1974). Levels of object representation in anaclitic and introjective depression. *The psychoanalytic study of the child*, *29*(1), 107-157.

Blatt, S., D'Afflitti & Quinlan, D. M. (1976). Experiences of Depression in Normal Young Adults. *Journal of Abnormal Psychology*, 85(4), 383 – 389.

Brenner, R. E., Heath, P. J., Vogel, D. L., & Credé, M. (2017). Two is more valid than one: Examining the factor structure of the Self-Compassion Scale (SCS). *Journal of Counseling Psychology.* DOI: 10.1037/cou0000211

Carver, C. S., & Ganellen, R. J. (1983). Depression and components of self-punitiveness: high standards, self-criticism, and overgeneralization. *Journal of abnormal Psychology*, *92*(3), 330.

Carver, C. S., La Voie, L., Kuhl, J., & Ganellen, R. J. (1988). Cognitive concomitants of depression: A further examination of the roles of generalization, high standards, and self-criticism. *Journal of Social and Clinical Psychology*, *7*(4), 350. http://dx.doi.org/10.1037/0021-843X.92.3.330

COSMIN – Systematic Reviews of Measurement properties (2016, January 1). Retrieved from:

http://www.cosmin.nl/Systematic%20reviews%20of%20measurement%20properties.html.

Cox, B. J., McWilliams, L. A., Enns, M. W., & Clara, I. P. (2004). Broad and specific personality dimensions associated with major depression in a nationally representative sample. *Comprehensive Psychiatry*, *45*(4), 246-253. http://dx.doi.org/10.1016/j.comppsych.2004.03.002

De Boer, M. R., Moll, A. C., De Vet, H. C., Terwee, C. B., Völker-Dieben, H. J., & Van Rens, G. H. (2004). Psychometric properties of vision-related quality of life questionnaires: a systematic review. *Ophthalmic and Physiological Optics*, *24*(4), 257-273. DOI: 10.1111/j.1475-1313.2004.00187

Fennig, S., Hadas, A., Itzhaky, L., Roe, D., Apter, A., & Shahar, G. (2008). Self-criticism is a key predictor of eating disorder dimensions among inpatient adolescent females. *International Journal of Eating Disorders,41*(8), 762-765. DOI: 10.1002/eat.20573

Furlan, A. D., Pennick, V., Bombardier, C., & van Tulder, M. (2009). 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine, 34*(18), 1929-1941. DOI: 10.1097/BRS.0b013e3181b1c99f

Gilbert, P (2005). *Compassion: Conceptualisations, Research and Use in Psychotherapy*. London: Routledge.

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in Psychiatric Treatment,* 15, 199-208. DOI: 10.1192/apt.bp.107.005264

Gilbert, P., Clarke, M., Hempel, S., Miles, J. N. V., & Irons, C. (2004). Criticizing and reassuring oneself: An exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology, 43*(1), 31-50. DOI: 10.1348/014466504772812959

Heinl, D., Prinsen, C.A.C., Deckert, S., Chalmers, J.R., Drucker, A.M., Ofenloch, R., Humphreys, R., Sach, T., Chamlin, S.L., Schmitt, J., Apfelbacher, C. Measurement properties of adult quality-of-life measurement instruments for eczema: a systematic review. Allergy 2016; 71: 358–370. DOI: 10.1111/all.12806

Ishiyama, F. I., & Munson, P. A. (1993). Development and validation of a self-critical cognition scale. *Psychological reports, 72*(1), 147-154. DOI: 10.2466/pr0.1993.72.1.147

Kamholz, B. W., Hayes, A. M., Carver, C. S., Gulliver, S. B., & Perlman, C. A. (2006). Identification and evaluation of cognitive affect-regulation strategies: Development of a self-report measure. *Cognitive Therapy and Research*, *30*(2), 227-262. DOI: 10.1007/s10608-006-9013-1

Kannan, D., & Levitt, H. M. (2013). A review of client self-criticism in psychotherapy. *Journal of Psychotherapy Integration*, *23*(2), 166. http://dx.doi.org/10.1037/a0032355

Kupeli, N., Chilcot, J., Schmidt, U. H., Campbell, I. C., & Troop, N. A. (2013). A confirmatory factor analysis and validation of the forms of self-criticism/reassurance scale. *British Journal of Clinical Psychology, 52*(1), 12-25. DOI: 10.1111/j.2044-8260.2012.02042.x

Luyten, P., Sabbe, B., Blatt, S. J., Meganck, S., Jansen, B., De Grave, C., Maes, F. & Corveleyn, J. (2007). Dependency and self-criticism: relationship with major depressive disorder, severity of depression, and clinical presentation. *Depression and anxiety*, 24, 586-596. DOI 10.1002/da.20272

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: challenges and benefits. *Research synthesis methods*,*5*(3), 221-234. DOI: 10.1002/jrsm.1106

Marshall, M. B., Zuroff, D. C., McBride, C., & Bagby, R. M. (2008). Self-criticism predicts differential response to treatment for major depression. *Journal of clinical psychology*, *64*(3), 231-244. DOI: 10.1002/jclp.20438

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81 – 90. DOI: 10.1037/0022-3514.52.1.81

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & De Vet, H. C. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*, *19*(4), 539-549. DOI: 10.1007/s11136-010-9606-8

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & de Vet, H. C. (2010b). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related

patient-reported outcomes. *Journal of clinical epidemiology*, *63*(7), 737-745. http://dx.doi.org/10.1016/j.jclinepi.2010.02.006

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Startford, P. W., Knol, D. L., Bouter, L. M., de Vet, H. CW. (2012). *COSMIN Checklist Manual.* Retrieved from: http://www.cosmin.nl/COSMIN%20checklist.html.

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and identity*, *2*(3), 223-250. DOI: 10.1080/15298860390209035

Neff, K. D., Whittaker, T. A., & Karl, A. (2017). Examining the factor structure of the Self-Compassion Scale in four distinct populations: Is the use of a total scale score justified?. *Journal of Personality Assessment*, 1-12. DOI: 10.1080/00223891.2016.1269334

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. Perspectives on Psychological Science, 3, 400-424.

Parker, G., Manicavasagar, V., Crawford, J. O., Tully, L., & Gladstone, G. (2006). Assessing personality traits associated with depression: the utility of a tiered model. *Psychological medicine*, *36*(08), 1131-1139. DOI: http://dx.doi.org/10.1017/S0033291706007562

Rector, N. A., Bagby, R. M., Segal, Z. V., Joffe, R. T., & Levitt, A. (2000). Self-criticism and dependency in depressed patients treated with cognitive therapy or pharmacotherapy. *Cognitive Therapy and Research*, *24*(5), 571-584. DOI: 10.1023/A:1005566112869

Santor, D. A., Zuroff, D. C., & Fielding, A. (1997). Analysis and revision of the Depressive Experiences Questionnaire: Examining scale performance as a function of scale

length. *Journal of Personality Assessment*, *69*(1), 145-163. http://dx.doi.org/10.1207/s15327752jpa6901_8

Schellingerhout, J. M., Verhagen, A. P., Heymans, M. W., Koes, B. W., Henrica, C., & Terwee, C. B. (2012). Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Quality of Life Research, 21*(4), 659-670. DOI: 10.1007/s11136-011-9965-9

Shahar, B., Carlin, E. R., Engle, D. E., Hegde, J., Szepsenwol, O., & Arkowitz, H. (2012). A pilot investigation of emotion-focused two-chair dialogue intervention for self-criticism. *Clinical Psychology & Psychotherapy*,*19*(6), 496-507. DOI: 10.1002/cpp.762

Shahar, B., Szsepsenwol, O., Zilcha-Mano, S., Haim, N., Zamir, O., Levi-Yeshuvi, S., & Levit-Binnun, N. (2015a). A Wait-List Randomized Controlled Trial of Loving-Kindness Meditation Programme for Self-Criticism. *Clinical psychology & psychotherapy*, *22*(4), 346-356. DOI: 10.1002/cpp.1893

Shahar, B., Doron, G., & Szepsenwol, O. (2015b). Childhood Maltreatment, Shame-Proneness and Self-Criticism in Social Anxiety Disorder: A Sequential Mediational Model. *Clinical psychology & psychotherapy*, *22*(6), 570-579. DOI: 10.1002/cpp.1918

Smart, L. M., Peters, J. R., & Baer, R. A. (2015). Development and validation of a measure of self-critical rumination. *Assessment*, 1073191115573300. DOI: 10.1177/1073191115573300

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., ... & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, *60*(1), 34-42. http://dx.doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*, *18*(8), 1115-1123. DOI: 10.1007/s11136-009-9528-5

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research*, *21*(4), 651-657. DOI: 10.1007/s11136-011-9960-1

Test-Retest Reliability Coefficient: Examples & Concept (2016, March 1). Retrieved from: http://study.com/academy/lesson/test-retest-reliability-coefficient-examples-lesson-quiz.html

Thompson, R., & Zuroff, D. C. (2004). The Levels of Self-Criticism Scale: comparative self-criticism and internalized self-criticism. *Personality and individual differences*, *36*(2), 419-430. http://dx.doi.org/10.1016/S0191-8869(03)00106-5

Van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board of the Cochrane Collaboration Back Review Group. (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine*, *28*(12), 1290-1299. DOI: 10.1097/01.BRS.0000065484.95996.AF

Verplanken, B. (2006). Beyond frequency: Habit as mental construct. *British Journal of Social Psychology*, *45*(3), 639-656. DOI: 10.1348/014466605X49122

Verplanken, B. &Aarts, H. (1999) Habit, Attitude, and Planned Behaviour: Is Habit an Empty Construct or an Interesting Case of Goal-Directed Automaticity? *European Review of Social Psychology*, 10, 101–134. DOI: 10.1080/14792779943000035

Verplanken, B., & Orbell, S. (2003). Reflections on Past Behavior: A Self-Report Index of Habit Strength1. *Journal of Applied Social Psychology*, *33*(6), 1313-1330. DOI: 10.1111/j.1559-1816.2003.tb01951.x

Verplanken, B., Friborg, O., Wang, C. E., Trafimow, D., & Woolf, K. (2007). Mental habits: Metacognitive reflection on negative self-thinking. *Journal of Personality and Social Psychology*, *92*(3), 526. DOI: 10.1037/0022-3514.92.3.526

Viglione Jr, D. J., Lovette, G. J., Gottlieb, R., & Friedberg, R. (1995). Depressive Experiences Questionnaire: An empirical exploration of the underlying theory. *Journal of personality assessment*, *65*(1), 91-99. http://dx.doi.org/10.1207/s15327752jpa6501_7

Weissman, A. N., & Beck, A. T. (1978). Development and validation of the Dysfunctional Attitude Scale: A preliminary investigation. Presented at the Annual Meeting of the American Educational Research Associatior, Toronto, Ontario, Canada.

Welkowitz, J., Lish, J. D., & Bond, R. N. (1985). The depressive experiences questionnaire: Revision and validation. *Journal of Personality Assessment*,*49*(1), 89-94. http://dx.doi.org/10.1207/s15327752jpa4901_17

Williams, M. J., Dalgleish, T., Karl, A., & Kuyken, W. (2014). Examining the factor structures of the five facet mindfulness questionnaire and the self-compassion scale. *Psychological Assessment*, *26*(2), 407. http://dx.doi.org/10.1037/a0035566

Figure 1 PRISMA Flow Chart: Original search and search update

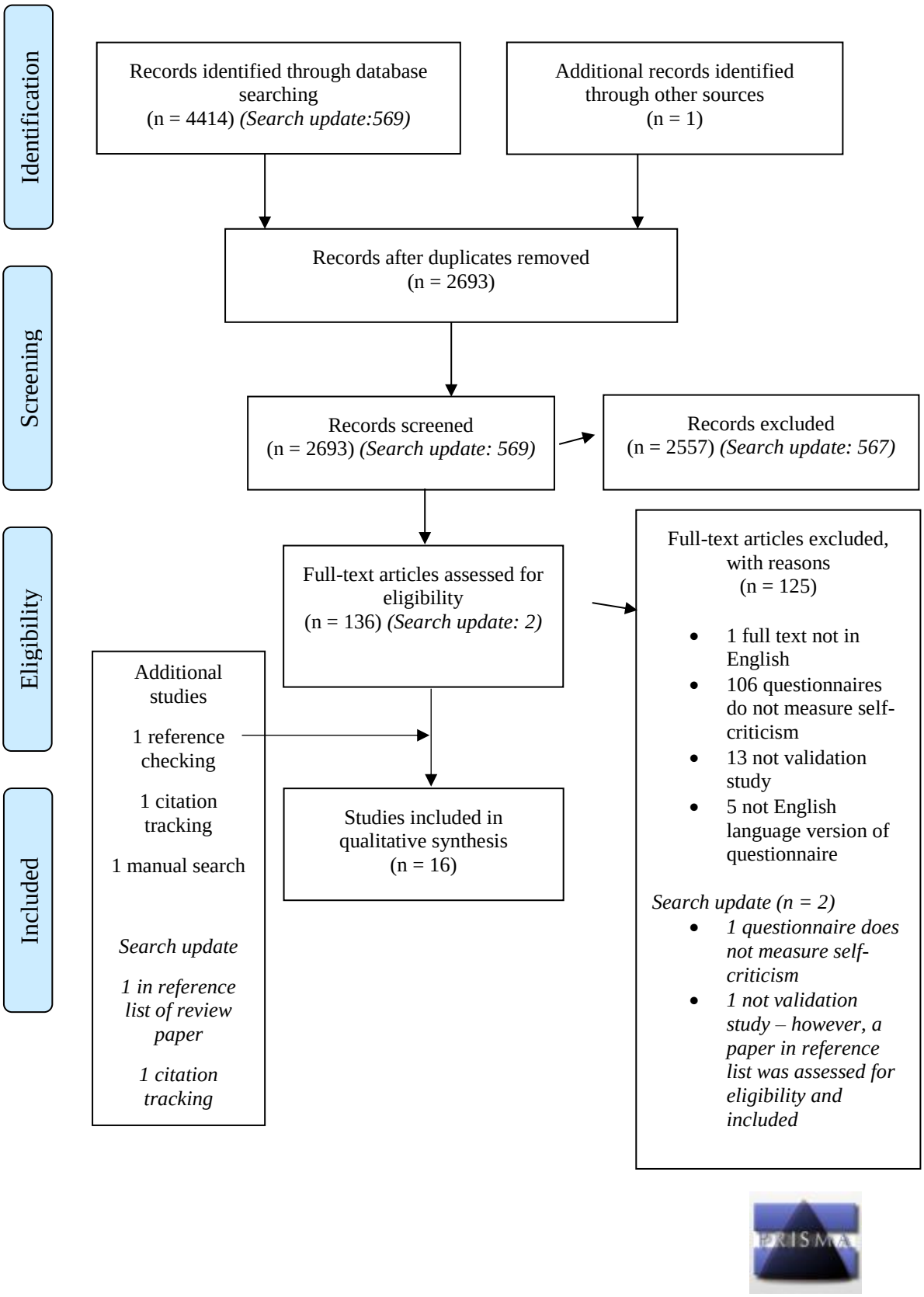

Table 1
Quality Criteria for Measurement Properties Assessed

Property	Definition	Quality Criteria Based On	Quality Criteria
Content Validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire	Terwee et al (2007)	+ A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection; ? A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method; - No target population involvement.
Internal Consistency	The extent to which items in a (sub) scale are intercorrelated, thus measuring the same construct	Terwee et al (2007)	+ Factor analyses performed on adequate sample size (7 * # items and ≥100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95; ? No factor analysis OR doubtful design or method; - Cronbach's alpha(s) <0.70 or >0.95, despite adequate design and method.
Reliability (Test-Retest)	The extent to which scores for participants who have not changed are the same for repeated measures over time	Terwee et al (2007)	+ ICC or weighted Kappa ≥ 0.70; ? Doubtful design or method (e.g., time interval not mentioned); - ICC or weighted Kappa < 0.70, despite adequate design and method.
Structural Validity	The degree to which the scores of a (sub) scale are an adequate reflection of the dimensionality of the construct to be measured	*Exploratory factor analysis* - Schellingerhout et al (2012)	+ Factors explain at least 50% of the variance ? Explained variance not mentioned - Factors explain <50% of the variance
		Confirmatory factor analysis – devised by authors	+ Factor structure confirmed ? Unclear if factor structure confirmed - Factor structure not confirmed
Construct Validity (Hypothesis Testing)	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured	Terwee et al (2007)	+ Specific hypotheses were formulated AND the majority of the results are in accordance with these hypotheses; ? Doubtful design or method (e.g., no hypotheses); - Less than 75% of hypotheses were confirmed, despite adequate design and methods.

Note. ICC: Intraclass correlation; + positive rating; ? indeterminate rating; - negative rating.

Table 2
Best Evidence Synthesis

Type of Self-Criticism Assessed	Original Authors	Questionnaire	Content Validity	Internal Consistency	Reliability	Structural Validity	Hypothesis Testing
Trait	Ishiyama & Munsun (1993)	Self-Critical Cognition Scale	Weak	+	? (limited)	+	Weak
Trait	Thompson & Zuroff (2004)	Levels of Self-Criticism Scale	Weak	+	Not studied	? (limited)	? (limited)
Trait (Subscale)	Carver & Ganellen (1983)	Attitudes Towards Self Scale	Weak	-	Not studied	-	Weak
Trait (Subscale)	Carver et al (1988)	Attitudes Towards Self Scale-Revised	Weak	-	? (limited)	+	Weak
Trait (Subscale)	Parker et al (2006)	Temperament & Personality Questionnaire	Weak	? (limited)	Weak	+	Not studied
Difficult situations	Gilbert et al (2004)	Forms of Self-Criticising/Attaching and Self-Reassuring Scale	Weak	++	Not studied	++	Conflicting findings
Difficult situations (Subscale)	Neff (2003)	Self-Compassion Scale	Weak	Conflicting findings	? (limited)	Conflicting findings	++
Mood regulation (Subscale)	Kamholz et al (2006)	Inventory of Cognitive Affect Regulation Strategies	Weak	Conflicting findings	Weak	Conflicting findings	Conflicting findings
Repetitive self-criticism	Verplanken et al (2007)	Habit Index of Negative Thinking	Weak	Weak	Weak	Not studied	++
Repetitive self-criticism	Smart, Peters & Baer (2015)	Self-Critical Rumination Scale	+++	+	? (limited)	++	+

Note.

Overall rating (i)	Level of evidence (ii)	Criteria (iii)
+++ ; ? (strong) ; ---	Strong	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality);
++ ; ? (moderate) ; - -	Moderate	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
+ ; ? (limited) ; -	Limited	One study of fair methodological quality
Conflicting findings	Conflicting	Conflicting findings across studies
Weak	Unknown	Only studies of poor methodological quality

(i) Direction of rating (positive, indeterminate or negative) was based on the measurement property ratings;
(ii) Level of evidence was based on the methodological quality of studies;
(iii) Criteria was adapted from Schellingerhout et al (2012) and Heinl et al (2016);

+ positive rating; ? indeterminate rating; - negative rating.

Table 3
Questionnaire Characteristics

Questionnaire	Type of Self-Criticism Assessed	Original Reference	Questionnaire Designed to Measure	Description of Items	Probe Statements & Example Items	Items for Scales or Self-Criticism Subscales	Response Options (Likert Scales)
Self-Critical Cognition Scale	Trait	Ishiyama & Munsun (1993)	To assess the "dispositional tendency" to process information about the self in a self-critical way. (P. 148). It has two subscales: 'Negative-self processing' & 'Failure in positive self-processing'.	Items focus on self-criticism, making negative social comparisons, an inability to keep a balanced perspective about oneself & exaggeration of negative aspects of oneself	Probe statement: unclear. *Negative self-processing: "*I tend to blow my weaknesses, limitations and mistakes out of proportion in my thinking"; *Failure in positive self-processing: "*I'm good at looking at myself critically while still remaining positive toward myself" (P. 150)	13	6-point (agree-disagree scale)
Levels of Self-Criticism Scale	Trait	Thompson & Zuroff (2004)	Self-criticism is conceptualised as a broad personality construct consisting of two developmental levels (Comparative Self-Criticism (CSC) and Internalised Self-Criticism (ISC)). CSC is defined as a negative view of oneself compared with other people. ISC is defined as a negative view of oneself compared with internalised personal standards.	No mention of self-criticism. CSC items focus on social anxiety, concerns & dilemmas. ISC items focus on affect & reactions to failure, high personal standards & experience of shame	Probe statement: unclear. *CSC:* "I don't spend much time worrying about what other people will think of me (Reversed)"; *ISC:* "When I don't succeed, I find myself wondering how worthwhile I am" (P. 424)	22	7-point (1 =not at all; 7= very well)

Questionnaire	Type of Self-Criticism Assessed	Original Reference	Questionnaire Designed to Measure	Description of Items	Probe Statements & Example Items	Items for Scales or Self-Criticism Subscales	Response Options (Likert Scales)
Attitudes Towards Self Scale	Trait (subscale)	Carver & Ganellen (1983)	The ATSS measures three self-regulatory vulnerabilities to depression (high-standards, overgeneralisation and self-criticism). **Self-criticism** is defined as making harsh judgements of oneself for failing to attain a standard.	No mention of self-criticism. Items focus on affect & reactions to failure & high personal standards	Probe statement: unclear. "When my behaviour doesn't live up to my standards, I feel I have let myself or someone else down" (P. 333)	4	5-point (1= extremely untrue, 5= extremely true)
Attitudes Towards Self Scale-Revised	Trait (subscale)	Carver et al (1988)	As above - the ATSSR was developed to produce “cleaner" subscales. (P. 352)	No mention of self-criticism. Items focus on affect & reactions to failure	Probe statement: unclear. "I get angry with myself if my efforts don't lead to the results I wanted" (P. 353)	3	5-point ('I agree very strongly' to 'I disagree very strongly', middle option 'neither agree nor disagree')
Temperament & Personality Questionnaire	Trait (subscale)	Parker et al (2006)	The TPQ measures personality traits or constructs thought to predispose individuals to depression. **Self-criticism** is defined as the tendency to be very tough on oneself.	Items focus on self-criticism, being tough/hard on oneself, high personal standards & a sense of satisfaction with oneself	Probe statement: "Please tick the option that best describes the way you usually or generally feel or behave (over the years and not just recently". Item: "I find it hard to measure up to my own standards" (available online)	Multiple versions of TPQ (81, 89 & 109-item versions). 4 or 8 in 109-item version	4-point (3= very true; 2= moderately true, 1= slightly true, 0= not true at all)

Questionnaire	Type of Self-Criticism Assessed	Original Reference	Questionnaire Designed to Measure	Description of Items	Probe Statements & Example Items	Items for Scales or Self-Criticism Subscales	Response Options (Likert Scales)
Forms of Self-Criticising/Attacking and Self-Reassuring Scale	In response to difficult situations	Gilbert et al (2004)	To assess forms of self-attacking when things go wrong for people. Separated into two forms: 'Inadequate self' focuses on attending to failures and inadequacies, and 'Hated self' focuses on more aggressive/disgust based self-attacking. Also, measures 'self-reassurance', defined as the ability to be reassuring to oneself when things go wrong.	Items focus on self-criticism, disliking oneself, not feeling good enough and other feelings about oneself associated with failure including disappointment, inadequacy, anger, frustration & disgust. There are also positively worded items about feeling good enough, loveable & acceptable	Probe statement: "When things go wrong for me…" Items: *Inadequate self:* "I remember and dwell on my failings"; *Hated self:*"I call myself names"; *Reassure self:*"I am able to remind myself of positive things about myself" (P. 37)	22 (Kupeli et al (2013) developed 18-item version)	5-point (0= not at all like me, 4= extremely like me)
Self-Compassion Scale	In response to difficult situations (subscale)	Neff (2003)	The SCS assesses levels of self-compassion in terms of 3 main components (divided into 6 subscales): self-kindness VS **self-judgement**; common humanity VS isolation; mindfulness VS over-identification. Self-judgement is conceptualised as a negative component of self-compassion, and is defined as being disapproving or judging of one's inadequacies and failures.	Items focus on self-judgement; being disapproving, intolerant & impatient about flaws, inadequacies & aspects of one's personality that you don't like	First probe statement: "How I typically act towards myself in difficult times" Item: "When I see aspects of myself that I don't like, I get down on myself" (available online)	5	5-point (1= Almost never, 5= Always always)

Questionnaire	Type of Self-Criticism Assessed	Original Reference	Questionnaire Designed to Measure	Description of Items	Probe Statements & Example Items	Items for Scales or Self-Criticism Subscales	Response Options (Likert Scales)
Inventory of Cognitive Affect Regulation Strategies	Mood regulation strategy (subscale)	Kamholz et al (2006)	The ICARUS assesses the deliberate and conscious cognitive affect-regulation strategies people use to reduce distressing emotions. **Self-criticism/self-blame** is defined as focusing on one's own perceived weakness and inadequacy.	Items focus on self-criticism, self-blame & thoughts about one's shortcomings, faults & mistakes. Also broader items focused on concentrating on negative emotions or repetitive thinking in response to negative emotions	First probe statement: "Indicate what you generally think about to make your mood better when you are sad..." Item: "I think about all my shortcomings, failings, faults and mistakes" (P. 231)	6	4-point (1= I don't do this at all; 2= I do this a little bit; 3= I do this a medium amount; 4= I do this a lot)
Habit Index of Negative Thinking	Repetitive self-criticism	Verplanken et al (2007)	A measure of the habit of negative self-thinking (adapted from the Self-Report Habit Index (Verplanken & Orbell, 2003). Focuses on the way a person thinks (as opposed to the content of thoughts).	Items focus on aspects of negative self-thoughts including whether they are frequent, automatic, unintentional & difficult to disengage from.	First probe statement: "Thinking negatively about myself is something…" Item: "I do frequently" (P. 541)	12	Verplanken (2006): 5-point (1= disagree completely, 5= agree completely). Verplanken et al (2007) used both 7-point & 5-point ('strongly disagree' to 'strongly agree')
Self-Critical Rumination Scale	Repetitive self-criticism	Smart, Peters & Baer (2015)	To assess self-critical rumination. Self-criticism is conceptualised as a form of negative thinking that focuses on devaluing oneself. Items also focus on ruminative qualities of thinking: "frequent, prolonged, repetitive & difficult to control". (P. 2).	Items focus on frequency & repetitiveness of self-criticism. Items also explore the content of thoughts, for example, whether someone focuses on aspects of themselves that they are ashamed of	First probe statement: unclear. "My attention is often focused on aspects of myself that I'm ashamed of" (P. 6)	10	4-point (1= not at all, 2=a little, 3= moderately, 4=very much)

Note. CSC: Comparative self-criticism; HS: Hated self; ICS: Internalised self-criticism; IS: Inadequate self; RS: Reassured self.

Table 4
Study Characteristics

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
SCCS	Ishiyama & Munsun (1993) Sample (1)	561	Undergraduates	N/A	22.3 (6.1) Victoria University; 27.1 (9.0) McGill university	Total sample - 210 males; 350 females;1 unidentified sex. Victoria University - 182 males; 272 females. McGill university 28 males; 78 females; 1 identified sex	Total sample = 40.3 (11.2); males = 39.6 (10.1); females = 40.8 (11.8).	Canada	NR
SCCS	Ishiyama & Munsun (1993) Sample (2)	142	Unclear	N/A	NR	83 males; 59 females	T1 = 39.1 (11.9). T2 = 38.3 (11.9)	Unclear	NR
LSCS	Thompson & Zuroff (2004) Study (1)	282	Undergraduates	N/A	NR	144 females; 138 males	N/A	USA/Canada	NR
LSCS	Thompson & Zuroff (2004) Study (2)	144	Undergraduates	N/A	NR	75 females; 69 males	NR	USA/Canada	NR
ATSS	Carver & Ganellen (1983) Sample (1)	1083	Undergraduates	N/A	NR	594 males; 489 females	N/A	USA	NR
ATSS	Carver & Ganellen (1983) Sample (2)	502	Undergraduates	N/A	NR	260 males; 242 females	See Table 6 - Construct validity	USA	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
ATSSR	Carver et al (1988) Study (1)	478	University students	N/A	NR	NR	N/A	USA	NR
ATSSR	Carver et al (1988) Study (2) & (4) (data combined)	Study 2 n = 170; Study 4 n = 219 (samples combined for analyses)	University students	N/A	NR	NR	NR	USA	NR
ATSSR	Carver et al (1988) Study (4) (subset of participants)	197	University students	N/A	NR	NR	NR	USA	NR
ATSSR	Carver et al (1988) Study (5) (depression group)	Depression group n = 5; Control group n = 11	Inpatients & hospital staff	Depression	NR	NR	NR	USA	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
ATSSR	Carver et al (1988) Study (5) (whole patient group)	70	Inpatients	24 Bipolar Disorder (12 in manic phase); 17 Schizophrenia; 7 SchizoAffective Disorder; 7 Atypical Psychosis; 5 Major Depression; 3 Dysthymic Disorder; 3 Adjustment Disorder; 2 Alcohol Dependence; 1 Schizophreniform Disorder; 1 Unspecified Nonpsychotic Mental Disorder	33.1 (9.43)	NR	NR	USA	NR
TPQ	Parker et al (2006) Sample (1)	529	Community sample (recruited at GP surgery)	N/A	35.5 (14.1)	54% females	N/A	Australia	NR
TPQ	Parker et al (2006) Sample (2)	52	Outpatients	Depression	41.3 (NR)	51.9% females	NR	Australia	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
FSCRS	Gilbert et al (2004)	246	Undergraduates	N/A	27.7 (7.2)	All females	Total sample - IS = 16.75 (8.44); HS = 3.86 (4.58); RS = 19.81 (5.92)	UK	NR
FSCRS	Kupeli et al (2013) Sample (1)	764	University students & community sample (recruited online)	N/A	28.6 (10.6)	Gender - 18.1% males (n = 138); 81.9% females (n = 626). Ethnicity - 76.2% White (n = 582).	N/A	UK	NR
FSCRS	Kupeli et al (2013) Sample (2)	806	As above	N/A	28.3 (10.6)	Gender - 17% males (n = 137); 83% females (n = 669). Ethnicity - 74.4% White (n = 600).	N/A	UK	NR
FSCRS	Kupeli et al (2013) Sample (3)	1224 (deduced by author)	Community sample (recruited online)	N/A	NR	NR	See Table 6 - Construct validity	UK	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
FSCRS	Baião et al (2015)	Non-clinical n = 887. Clinical n = 171 (after 4 excluded)	Secondary analyses on data from 12 previous studies (7 non-clinical; 5 clinical groups)	100 (58.48%) Depression; 16 (9.36%) Personality Disorder; 13 (7.60%) Substance Abuse; 9 (5.26%) Anxiety; 3 (1.54%) Bipolar Disorder. (Missing data = 30)	Non-clinical population = 24.13 (7.79). Clinical population = 44.22 (12.05) (missing data = 23 clinical participants)	Non clinical population - 210 males; 676 females. Clinical population - 67 males; 91 females (missing data for 13 clinical participants)	See Table 6 - Construct validity	UK	NR
SCS	Neff (2003) Study (1) - content validity	Focus group n = 68. Piloting of items n = 71	Undergraduates	N/A	Focus group= 21.7 (2.32). Piloting of items = 21. (2.03)	Focus group - 30 males; 38 females. Piloting of items - 24 males; 47 females	N/A	USA	N/A
SCS	Neff (2003) Study (1) - main study	391	Undergraduates	N/A	20.91 (2.27)	Gender - 166 males; 22 females. Ethnicity - 58% White; 21% Asian; 11% Hispanic; 4% Black; 6% other	Total sample = 3.14 (0.79); Males = 3.00 (0.81); Females = 3.24 (0.77)	USA	NR
SCS	Neff (2003) Study (2)	232	Undergraduates	N/A	21.31 (3.17)	Gender - 87 males; 145 females. Ethnicity - 58% White; 22% Asian; 14% Hispanic; 3% Black; 3% other.	NR	USA	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
SCS	Neff (2003) Study (3)	Students n = 232; Buddhist n = 43	Community sample (recruited from Buddhist email list subscription)	N/A	Students = 21.31 (3.17); Buddhists = 47 (9.71)	Students: Gender - 87 males; 145 females. Ethnicity - 58% White; 22% Asian; 14% Hispanic; 3% Black; 3% other. Buddhists: Gender - 16 males; 27 females. Ethnicity - 91% White; 5% Asian; 2% other.	See Table 6 - Construct validity	USA	NR
SCS	Williams et al (2014) Sample (1)	821	Community sample (recruited online)	N/A	25.7 (9.8)	Gender - 697 females (74.1%). Ethnicity - 800 (85.1%) White; 140 14.9%) Other	12.10 (4.40)	UK	EX
SCS	Williams et al (2014) Sample (2)	211	Community sample (recruited online)	N/A	46.51 (13.1)	Gender - 153 females (65.1%). Ethnicity - 216 (91.9%) White; 19 (8.1%) Other	17.15 (4.29)	UK	EX
SCS	Williams et al (2014) Sample (3)	390	Community sample (recruited through MBCT trial)	Recurrent Depressive Disorder	50.16 (11.8)	Gender - 325 females (76.6%). Ethnicity - 410 (96.7%) White; 4 (0.9%) Other; 10 (2.4%) Missing	11.81 (3.93)	UK	EX

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
SCS	Neff, Whittaker & Karl (2017) Sample (1)	222	Undergraduates	N/A	20.94 (2.03)	Gender - 84 males; 138 females. Ethnicity - 57% White; 22% Asian; 14% Hispanic; 3% Black; 4% other	3.00 (0.81)	USA	NR
SCS	Neff, Whittaker & Karl (2017) Sample (2)	1394	Community sample (recruited online)	N/A	36.01 (12.88)	Gender – 35% males; 65% females. Ethnicity - 77% White; 6% Asian; 6% Hispanic; 7% Black; 6% other	3.11 (0.96)	USA	NR
SCS	Neff, Whittaker & Karl (2017) Sample (3)	215	Meditators (recruited online)	N/A	47.40 (11.59)	Gender – 30% males; 70% females. Ethnicity - 87% White; 2% Asian; 2% Hispanic; 2% Black; 7% other	2.64 (0.78)	USA	NR
SCS	Neff, Whittaker & Karl (2017) Sample (4)	390	Clinical (reanalysis of Williams et al, 2014, Sample 3)	Recurrent Depressive Disorder	50.16 (11.8)	Gender – 23% males; 77% females	3.64 (0.78)	UK	EX
SCS	Brenner, Heath, Vogel & Credé (2017)	1115	Undergraduates	N/A	19.4 (1.7)	Gender – 56.4% females; 43.6 males. Ethnicity – 78.1% European American; 9.1% Asian American/Pacific Islander; 2.8% Latino;	3.33 (0.83)	USA	EX

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
						4.4% African American; 0.4% American Indian/Alaskan Native; 0.1% Native Hawaiian; 3.3% Multi-Racial; 1.0% Self-Identify			
ICARUS	Kamholz et al (2006) Study (1)	Pilot study 1 n = 193; Pilot study 2 & main sample used n = 398 (after 28 excluded)	Undergraduates	N/A	86.2% = 21 years and younger	Gender - 59% females. Ethnicity - 44.8% Caucasian; 30.7% Hispanic; 12.1% African American; 7.1% Asian; 5.3% Other or mixed	2.38 (0.64)	USA	EX
ICARUS	Kamholz et al (2006) Study (2A)	132	Undergraduates	N/A	20.27 (3.24)	Gender 62% females. Ethnicity - 46.8% Caucasian; 34.1% Hispanic; 10.3% African American; 8.7% Asian	1.49 (0.61)	USA	NR
ICARUS	Kamholz et al (2006) Study (2B)	132	Undergraduates	N/A	As above	As above	As above	USA	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
ICARUS	Kamholz et al (2006) Study (3)	208	Outpatients	137 (66%) Substance-Use Disorder; 129 (62%) at least one Axis I psychiatric diagnosis; 93 (45%) two diagnoses; 62 (30%) three diagnoses. 91 (71%) mood disorder; 54 (42%) PTSD; 33 (26%) non-PTSD anxiety disorder 48; (37%) Psychotic Disorder	49 (7.98)	Gender - n = 201, 97% males. Ethnicity - n = 140, 67.3% Caucasian; n = 54, 26% African-American; n = 2, 1% Hispanic, n = 4, 1.9% Native American; 8, 3.8% Other.	2.38 (0.75)	USA	NR
HINT	Verplanken (2006) Study (2)	194	University students	N/A	NR	123 females;71 males	2.32 (1)	Norway	NR
HINT	Verplanken et al (2007) Study (1)	157	University students	N/A	NR	95 females; 61 males (1 participant did not disclose)	NR	Norway	NR
HINT	Verplanken et al (2007) Study (4)	155	University students	N/A	NR	88 females; 66 males (1 participant did not disclose)	2.70 (1.05)	Norway	NR
HINT	Verplanken et al (2007) Study (5)	125	University students	N/A	NR	79 females; 46 males	3.03 (1.36)	USA	NR

Questionnaire	Author(s)	N	Population	Diagnoses	Age - Mean (SD)	Demographic Information	Means & SDs (Scales or Self-Criticism Subscales)	Country	Missing Items
HINT	Verplanken et al (2007) Study (8)	T1: n = 1682. T2: n = 1102	Community sample (recruited via postal system)	N/A	40.27 (8.23)	T1: 939 females; 736 males (7 did not disclose). T2: 641 females; 461 males	T1 = 2.72 (1.56). T2: NR	Norway	EX
SCRS	Smart, Peters & Baer (2015) Study (1)	Undergraduates n = 25; adult outpatient n = 13	Undergraduates & outpatients	N/A	NR	NR	N/A	USA	N/A
SCRS	Smart, Peters & Baer (2015) Study (2)	420 (after 90 excluded)	Undergraduates	N/A	18.99 (1.44)	Gender - 51.9% females. Ethnicity - 71.9% Caucasian.	2.17 (0.73)	USA	NR
SCRS	Smart, Peters & Baer (2015) Study (3)	143	Undergraduates	N/A	19.00 (1.46)	Gender - 69.9% females. Ethnicity - 72.2% Caucasian.	N/A	USA	NR
SCRS	Smart, Peters & Baer (2015) Study (4)	70	Undergraduates	N/A	NR	Gender - 89.9% female. Ethnicity - 91.3% Caucasian	T1 = 1.90 (SE = 0.08); T2 = 1.83 (SE = 0.08)	USA	NR

Note. ATSS: Attitudes Towards Self Scale; ATSR: Attitudes Towards Self Scale-Revised; FSCRS: Forms of Self-Criticising/Attaching and Self-Reassuring Scale; HINT: Habit Index of Negative Thinking; HS: Hated self; ICARUS: Inventory of Cognitive Affect Regulation Strategies; IS: Inadequate self; LSCS: Levels of Self-Criticism Scale; N/A: Not applicable; NR: Not recorded; RS: Reassured self; SD: standard deviation; SCCS: Self-Critical Cognition Scale; SCRS: Self-Critical Rumination Scale; SCS: Self-Compassion Scale; TPQ: Temperament & Personality Questionnaire; T1: Time 1; T2: Time

Table 5
Ratings for Methodological Quality and Measurement Properties

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
SCCS	Ishiyama & Munsun (1993) Sample (1)	Poor	?	Fair	[+ Negative self-processing: 0.89; Failure in positive self-processing: 0.77]			Fair	[+ NSP = 43%; FPSP = 9.1%]
SCCS	Ishiyama & Munsun (1993) Sample (2)					Fair	[? $ Test-retest reliability - r138 = 0.83 for total sample; r81 = 0.82 for males; r57 = 0.86 for females. TI: 6.5 weeks]		
LSCS	Thompson & Zuroff (2004) Study (1)	Poor	?	Fair	[+ CSC 0.81; ISC 0.87]			Fair	[?]
ATSS	Carver & Ganellen (1983) Sample (1)	Poor	?	Fair	[- Self-criticism: 0.65]			Fair	[- 40%]
ATSSR	Carver et al (1988) Study (1)	Poor	?	Fair	[- Self-criticism: 0.65]			Fair	[CFA +]

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
ATSSR	Carver et al (1988) Study (2) & (4) (data combined)					Fair	[? $ Test-retest correlations - Self-criticism: 0.59.TI: 6 weeks]		
TPQ	Parker et al (2006) Sample (1)	Poor	?	Fair	[? Cronbach alphas - ranged from 0.62 to 0.91 (Individual subscales not reported)]			Fair	[+50%]
TPQ	Parker et al (2006) Sample (2)					Poor	[? ICCs recorded for each subscale - Self-criticism: 0.73 (p<0.001). TI: mean = 29 days (range 5 - 150 days)]		
FSCRS	Gilbert et al (2004)	Poor	?	Fair	[+ IS: 0.90; RS: 0.86; HS: 0.86]			Fair	[+ 58.32%]
FSCRS	Kupeli et al (2013) Sample (1)							Fair (Items 4, 18 & 20 removed due to low factor loadings)	[+ IS = 47.52; HS = 8.8%; RS = 6.74%]

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
FSCRS	Kupeli et al (2013) Sample (2)			Fair	[+ 18-items - IS: 0.90; RS: 0.88; HS: 0.83. Original 22-items - IS: 0.91; RS: 0.88; HS: 0.86]			Fair (Items 4, 18, 20 & 22 removed due to low factor loadings)	[CFA +]
FSCRS	Baião et al (2015)			Fair	[+ Non-clinical - IS: 0.90; RS: 0.85; HS: 0.85. Clinical - IS: 0.91; RS: 0.85; HS: 0.87]			Fair	[CFA +]
SCS	Neff (2003) Study (1) - content validity	Poor	?						
SCS	Neff (2003) Study (1) - main study			Fair	[+ Self-judgement = 0.77]			Fair	[?]
SCS	Neff (2003) Study (2)					Fair	[? $ Test-retest reliability - Self-judgement = 0.88. TI: 3 weeks]	Fair	[?]
SCS	Williams et al (2014) Sample (1)			Fair	[? Self-judgement = 0.8]			Excellent	[CFA -]

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
SCS	Williams et al (2014) Sample (2)			Fair	[? Self-judgement = 0.82]			Excellent	[CFA -]
SCS	Williams et al (2014) Sample (3)			Fair	[? Self-judgement = 0.78]			Good	[CFA -]
SCS	Neff, Whittaker & Karl (2017) Sample (1)			Poor	[? Self-judgement = 0.85. (No FA)]			Fair	[CFA +]
SCS	Neff, Whittaker & Karl (2017) Sample (2)			Poor	[? Self-judgement = 0.85. (No FA)]			Fair	[CFA +]
SCS	Neff, Whittaker & Karl (2017) Sample (3)			Poor	[? Self-judgement = 0.89. (No FA)]			Fair	[CFA +]

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
SCS	Neff, Whittaker & Karl (2017) Sample (4)			Poor	[? Self-judgement = 0.78. (No FA)]			Good	[CFA -]
SCS	Brenner, Heath, Vogel & Credé (2017)			Poor	[? Total SCS = 0.93 (Self-judgement subscale not recorded) (No FA)]			Good	[CFA ?]
ICARUS	Kamholz et al (2006) Study (1)	Poor	-	Good	[? Self-Criticism/Self-Blame = 0.81]			Good	[?]
ICARUS	Kamholz et al (2006) Study (2A)			Good	[? Self-criticism/self-blame = 0.83]				
ICARUS	Kamholz et al (2006) Study (3)			Poor	[? Self-criticism/self-blame = 0.85]	Poor	[? $ Test-retest reliability correlation coefficients - self-criticism/self-	Poor	[+73.5%]

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
							blame = 0.65 (p<0.001). TI: 1 month]		
HINT	Verplanken (2006) Study (2)			Poor	[? 0.95. (No FA)]				
HINT	Verplanken et al (2007) Study (1)	Poor	?	Poor	[? 0.943. (No FA)]				
HINT	Verplanken et al (2007) Study (4)			Poor	[? 0.945. (No FA)]				
HINT	Verplanken et al (2007) Study (5)			Poor	[? 0.947. (No FA)]				
HINT	Verplanken et al (2007) Study (8)			Poor	[? 0.955. (No FA)]	Poor	[? $ Test -retest reliability = 0.801 (p<0.01). TI: 9 months]		
SCRS	Smart, Peters & Baer (2015) Study (1)	Excellent	+						

	Author(s)	Content Validity	Content Validity	Internal Consistency	Internal Consistency (i)	Reliability	Reliability	Structural Validity	Structural Validity (ii)
		Methodological Quality	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*	*Methodological Quality*	*Measurement Property*
SCRS	Smart, Peters & Baer (2015) Study (2)			Fair	[+ 0.92.]			Fair	[+ 58.4%]
SCRS	Smart, Peters & Baer (2015) Study (3)							Fair	[CFA +]
SCRS	Smart, Peters & Baer (2015) Study (4)					Fair	[? $ Test-retest correlation = 0.86 (& no statistical difference found between scores). TI: 13 -37 days]		

Note. ATSS: Attitudes Towards Self Scale; ATSR: Attitudes Towards Self Scale-Revised; CSC: Comparative self-criticism; FSCRS: Forms of Self-Criticising/Attaching and Self-Reassuring Scale; HINT: Habit Index of Negative Thinking; HS: Hated self; ICARUS: Inventory of Cognitive Affect Regulation Strategies; ICS: Internalised self-criticism; IS: Inadequate self; LSCS: Levels of Self-Criticism Scale; RS: Reassured self; SCCS: Self-Critical Cognition Scale; SCRS: Self-Critical Rumination Scale; SCS: Self-Compassion Scale; TPQ: Temperament & Personality Questionnaire; CFA: Confirmatory factor analysis; ICC: Intraclass correlation coefficient; TI: Time interval; T1: Time 1; T2: Time 2; $: Statistical test other than what COSMIN recommends; i:Cronbach's alpha presented; ii: percentage of variance explained presented.

Table 6
Construct Validity - Ratings for Methodological Quality and Measurement Property

Questionnaire	Author(s)	Methodological Quality	Measurement Property (i)	Results for Scales or Self-Criticism Subscales
SCCS	Ishiyama & Munsun (1993) Sample (1)	Poor	?	Correlation coefficients (p<0.05) 1. Self-esteem (n = 416) = -0.71. 2. Depression (n = 168) = 0.42. 3. Between group comparison: higher count of negative self-descriptive adjectives in 'high' self-critical group (p<0.01).
LSCS	Thompson & Zuroff (2004) Study (2)	Fair	?	Correlation coefficients (P<0.05) for CSC :1. Distress = 0.53; 2. Self-esteem = -0.66;3. Perfectionism-self = 0.21; Perfectionism-other = 0.21; Perfectionism-social = 0.46. For ISC: 1. Distress = 0.44. 2. Self-esteem = -0.52.3. Perfectionism-self = 0.45. Perfectionism-other = 0.24. Perfectionism-social = 0.49.
ATSS	Carver & Ganellen (1983) Sample (2)	Poor	?	Between group comparison (p<0.02): Gender - Males: Mean = 15.08 (SD = 3.19); Females: Mean = 15.79 (SD = 3.43)
ATSSR	Carver et al (1988) Study (2) & (4) (data combined)	Poor	?	Correlation coefficients (*p<0.05 **p<0.01) Depression (study sample 2) = 0.15*; (study sample 4) = 0.26**.
ATSSR	Carver et al (1988) Study (4) (subset of participants)	Poor	?	No results presented
ATSSR	Carver et al (1988) Study (5) (depression group)	Poor	?	No results presented
ATSSR	Carver et al (1988) Study (5) (whole patient group)	Poor	?	No results presented

Questionnaire	Author(s)	Methodological Quality	Measurement Property (i)	Results for Scales or Self-Criticism Subscales
FSCRS	Gilbert et al (2004)	Fair	+	Correlation coefficients (* = <0.05; ** =<0.01) for IS: 1. Depression = 0.52*; 2. ISC = -0.77**. CSC = 0.63**. For HS: 1. Depression = 0.57**. 2. ISC = 0.45**. CSC = 0.55**. For RS: 1. Depression = -0.51**. 2. ISC = -0.45**. CSC = -0.63**
FSCRS	Kupeli et al (2013) Sample (3)	Fair	+	Correlation coefficients (**p<0.001) 18-item FSCRS 1. Happiness - RS = -0.66**; HS = -0.66**; IS = -0.60**. 22-item FSCRS 1. Happiness - RS = -0.66**; HS = -0.66**; IS = -0.62**. 2. Between group comparison with 18-item version: Gender - females - **IS Mean = 18.3 (SD = 6.4); **RS Mean = 22.2 (SD = 6.8); HS Mean = 9.0 (SD = 4.9). Males - **IS Mean = 16.3 (SD = 6.5); **RS Mean = 20.6 (SD = 7.0); HS Mean = 8.5 (SD = 4.4)
		Fair	?	1. Between group comparison (**p = 0.000): **RS - Clinical: Mean = 10.68 (SD = 6.51); Non-clinical: Mean = 20.27 (SD = 5.77). **IS Clinical: Mean = 27.47 (SD = 7.51); Non-clinical: Mean = 17.72 (SD = 8.29). **HS Clinical: Mean = 12.26 (SD = 5.67); Non-clinical: Mean = 3.88 (SD = 4.59). 2. Between group comparison: No significant differences found for gender in clinical population. Gender in non-clinical population - **RS - males: Mean = 21.20 (SD = 5.27); females: Mean = 19.98 (SD = 5.90) **IS - males: Mean = 16.42 (SD = 7.44); females: Mean = 18.11 (SD = 8.50). **HS - males: Mean = 3.36 (SD = 3.71); females Mean = 4.05 (SD = 4.83) (p = .058)
FSCRS	Baião et al (2015)			
SCS	Neff (2003) Study (1) - main study	Fair	+	Between group comparison (p<0.005) - Gender - Males Mean = 3.00 (SD = 0.81). Females Mean = 3.24 (SD = 0.77)
SCS	Neff (2003) Study (2)	Fair	+	No results presented
SCS	Neff (2003) Study (3)	Fair	+	Between group comparison (p<0.001) - Buddhist Mean = 2.20 (SD = 0.65); Students Mean 3.07 (SD = 0.82)
ICARUS	Kamholz et al (2006) Study (1)	Poor	?	Between group comparison - Gender - 2x15 (gender by strategy) repeated measures ANOVA completed. No significant interactions found.

Questionnaire	Author(s)	Methodological Quality	Measurement Property (i)	Results for Scales or Self-Criticism Subscales
ICARUS	Kamholz et al (2006) Study (2A)	Fair	?	Mood induction experiment to test predictive validity - no correlations presented.
ICARUS	Kamholz et al (2006) Study (2B)	Poor	?	Between group comparison - Gender - 2x15 (gender by strategy) repeated measures ANOVA completed. No significant interactions found.
ICARUS	Kamholz et al (2006) Study (3)	Fair	+	Correlation coefficients (*p<0.05 **p<0.01 ***p<0.001) 1. Depression = 0.60***; 2. Anxiety = 0.57***
HINT	Verplanken (2006) Study (2)	Fair	+	Correlation coefficients (**P<0.001) 1. Past frequency of 'negative self-thinking' = 0.648**. 2. Self esteem = -0.737**. 3. Depressive/anxiety symptoms = 0.571**.
HINT	Verplanken et al (2007) Study (1)	Fair	+	Task used to test hypotheses (Story & thought-listing protocol). HINT correlated significantly with negative self-thoughts (r = 0.295, p<0.001). Correlation between HINT and negative self-thoughts was significantly different to HINT and general negative thoughts (z = 2.02, p<0.05).
HINT	Verplanken et al (2007) Study (4)	Fair	+	Correlation coefficients (p<0.001) 1. Rumination = 0.665; 2. Self-esteem = -0.555
HINT	Verplanken et al (2007) Study (5)	Fair	+	Corelation coefficients (**p<0.01, ***p<0.001) 1. 'Negative self-thinking' = 0.537***; 2. Explicit self-esteem = -0.473***; 3. Implicit self-esteem = -0.279**
SCRS	Smart, Peters & Baer (2015) Study (2)	Fair	+	Correlation coefficients (*p<0.05; **p<0.01) 1. Rumination = 0.81**; 2. Brooding = 0.68**. 3. Rumination-anger = 0.67**; 4. Rumination-anxiety = 0.59**; 5. Rumination-interpersonal = 0.53**; 6. Rumination-social situations = 0.65**; 7. Self-criticism = 0.81**; 8. Shame (different measures) = 0.55**; 0.66**; 0.73**; 9. Self-compassion = -0.62**; 10. Depression/anxiety = 0.58**

Note. ATSS: Attitudes Towards Self Scale; ATSR: Attitudes Towards Self Scale-Revised; CSC: Comparative self-criticism; FSCRS: Forms of Self-Criticising/Attaching and Self-Reassuring Scale; HINT: Habit Index of Negative Thinking; HS: Hated self; ICARUS: Inventory of Cognitive Affect Regulation Strategies; ICS: Internalised self-criticism; IS: Inadequate self; LSCS: Levels of Self-Criticism Scale; RS: Reassured self; SCCS: Self-Critical Cognition Scale; SCRS: Self-Critical Rumination Scale; SCS: Self-Compassion Scale; TPQ: Temperament & Personality Questionnaire.

(i) + Specific hypotheses were formulated AND the majority of the results are in accordance with these hypotheses; ? Doubtful design or method (e.g., no hypotheses); - Less than 75% of hypotheses were confirmed, despite adequate design and methods.